1. REPORT DATE (dd-mm-yy) February 2001	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) 1995-98

4. TITLE AND SUBTITLE Issues of Adaptive Automated Surveys in a Computer Network Environment	5a. CONTRACT OR GRANT NUMBER DASW01-97-C-0011
	5b. PROGRAM ELEMENT NUMBER 433709

6. AUTHOR(S) Elliot E. Entin, Caroline Kerrigan, Michael Berbaum, Patrice Lancey, & Debra McCallum (ALPHATECH, Inc.)	5c. PROJECT NUMBER
	5d. TASK NUMBER 7001
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) ALPHATECH, Inc. 50 Mall Road Burlington, MA 01803-4562	8. PERFORMING ORGANIZATION REPORT NUMBER TR-830

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Study Report 2001-02

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

COR: Morris Peterson

14. ABSTRACT *(Maximum 200 words)*: Computer-based surveys administered over a computer network hold out a myriad of possibilities for tailoring surveys to particular groups or even to individual respondents, including adaptive automated surveys and new types of surveys never possible before. Computer-administered surveys allow survey items and instructions to be conveyed textually, graphically, and even in animation. Computer networks thus provide an ideal medium on which to conduct innovative, multimedia, dynamic surveys--surveys that can be sent instantaneously to a large number of recipients, regardless of whether the recipients are on-line at that time. They also allow the recipients to reply at a time that is convenient for them, without regard to whether the originator of the questionnaire is currently on-line and without having to look up an address or find a post box. The goals for this research included a review of the state of the art in current survey technology, an analysis of which methods and procedures can be applied directly to computer network surveys, and hypothesized extrapolations of certain aspects of current survey technology that hold promise for the new medium of computer networks. Based on these findings, we delineated four sets of critical issues that must be investigated in order to conduct valid and reliable network surveys. We also conducted two pilot experiments to explore initial hypotheses about how effective network questionnaires should be formatted and how supporting help could be offered. These pilot experiments provided an opportunity to test and validate the experimental design and methodology that we developed. The appendix of this report discusses four areas in which empirical research is needed and outlines a program for conducting such research.

15. SUBJECT TERMS

survey methodology, computer-based surveys, adaptive automated surveys, computer networks

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number) Morris Peterson 703-617-7803
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	54	

ISSUES OF ADAPTIVE AUTOMATED SURVEYS IN A COMPUTER NETWORK ENVIRONMENT

August 1997

Prepared by

Elliot E. Entin[•]
Caroline Kerrigan
Michael Berbaum
Patrice Lancey
Debra McCallum

[•] Now at Aptima, Inc., 600 W. Cummings Pk., Woburn, MA 01801

A primary mission of the Army Personnel Survey Office (APSO) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is to collect information on a wide range of issues important to the Army. These findings provide the Army with timely information on which to base future planning and policy making.

This Study Report addresses the topic of automated adaptive surveys. Computer-based surveys administered over a computer network have the potential for tailoring surveys to particular groups or even to particular individuals. This report surveys the literature on the subject, discusses the critical issues involved in automating surveys, and describes the results of a pilot test developed to test the conclusions derived from this research.

The Army can use the findings of this report to assist its survey collection efforts.

EDGAR M. JOHNSON
Director

ISSUES OF ADAPTIVE AUTOMATED SURFEYS IN A COMPUTER NETWORK ENVIRONMENT

EXECUTIVE SUMMARY_____

Research Requirement:

       The Army employs surveys to collect information on a wide range of important issues. As in many other areas, survey technology changes. Hence this research effort was geared to identify what we already know about survey technology and methodology, what we can generalize from what is known, and what new knowledge we need to develop a complete methodology to conduct effective surveys over a computer network.

Procedure:

       We reviewed and analyzed research on existing survey technologies, summarized the state of empirical knowledge with respect to principles and procedures of survey instrument construction and administration, and identified issues specific to the conduct of surveys for computer networks. We further designed and implemented two pilot experiments to investigate response format effects and graphical user interfaces.

Findings:

       We found that the applicability of empirical findings to automated network surveys ranged from very high to very low. Principles, procedures, and practices were delineated that are applicable to network surveys and ready to use—knowledge and procedures we can take as firmly established by previous research. For example, a great deal of the accumulated knowledge about question wording should be perfectly applicable in network surveys, as should knowledge about question ordering (as when one question evokes an evaluatively loaded cultural frame of reference that then influences responses to a second question).

       The results of the two pilot experiments indicated that textually based enhancements and encouragements were capable of producing almost error-free responses and that the use of certain Guided User Interfaces (GUIs) could significantly increase the reliability of the response data. The experimental design and methodology of the two pilot experiments proved effective and provided a valid prototype for the design and methodology of future Phase II studies. The use of two different populations, college students and older individuals in the Army Reserve demonstrated the robustness of the prototype methodology and results. In Appendix A we have identified four important areas for future research and suggested a general blueprint for the experimental designs. Each proposed area of study would fill the gaps in our knowledge and help us to develop a more complete methodology for network surveys.

<u>Utilization of Findings</u>:

Computer networks such as the Internet and the World Wide Web hold great potential for information gathering and research of all kinds. The research and development described in this report provide a reliable and valid methodology for the Army to employ in order to reap the benefits of using a computer network to conduct surveys.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Issues of Adaptive Automated Surveys in a Computer Network Environment

INTRODUCTION

<u>Background and Scope</u>

Surveys have been used to examine a myriad of topics ranging from very private concerns of the individual to their experiences with a variety of consumer products (Rossi et al., 1983). These researchers note that there appears to be no bound on the kinds of questions that can be asked in a survey, nor does there appear to be a limit to the willingness of individuals to take the time to complete them. Individuals seem to be particularly enthusiastic about responding to surveys delivered over a computer network. Walsh, Kiesler, Sproull, and Hesse (1992) noted that while conducting a survey of 300 oceanographers over a computer network, an additional 104 individuals spontaneously asked to participate. Moreover, participants in this self-selected sample were among the first to completed the 93-item, 30-minute survey and personally bore the estimated $5.00 cost in net charges to do so.   One concern brought about by self-selected respondents is that the sample may be very strongly biased towards those with access to computer networks (Parker, 1992).

The current effort is based on two basic premises. The first is that computer networks hold great promise for the field of survey research. Network-administered surveys can be sent out almost instantaneously to a huge, diverse sample. There is no need for a simultaneous connection, as in a telephone or face-to-face survey. Responses can be obtained rapidly-- whenever the respondents are logged on to the network. There is no need to wait until the respondents remember to put the survey in a postbox, as in mail surveys.

More significantly, the application of computer-based surveys administered over a network holds out a myriad of possibilities for tailoring surveys to particular groups or even to individual respondents, including adaptive automated surveys and new types of surveys never possible before. Surveys no longer need to be static. Rather, surveys implemented and administered on a computer can take advantage of the computer's ability to monitor the respondents (making it possible to present questions and question sequences in an adaptive manner, prompt, and offer help), maintain quality control, prepare analyses of respondents' answers, and implement new survey procedures that were heretofore not possible. For example, modularly designed surveys might be sent out via a network to various sites where the *client's* computer (executing a Java applet) parcels out the survey parts to individuals possessing unique characteristics or information for completion. Then the computer reassembles the parts, analyzes, and sends the results back over the network to the *host*. Animation, as well as high quality pictures and graphics, can be folded into a survey to increase participation, simplify instructions, illustrate a process or entity to be evaluated, and serve a myriad of other purposes.

There are several challenges common to any survey, no matter how it is administered. One is the identification of the target population of interest and, if the survey is not to be administered to the entire population, the selection of a representative sample from that

population. Once the target sample has been selected, the challenges are to develop appropriate and clearly worded questions, to maximize the response rate, and to obtain completed surveys, with thorough, honest, and stable answers to each question. In other words, the goal is to obtain a valid and reliable survey, where validity concerns whether respondents are representative of the overall population and whether the survey questions get at the underlying issues being explored in the survey, and where reliability concerns obtaining stable and accurate answers to the questions.

The second premise of this work is that a sizable and mature literature on survey technology exists and that it would be foolhardy not to take advantage of this firm foundation of guidelines for reliable and valid surveys. The most efficient way to develop new technology is to build upon what is already known. Many issues that must be addressed when evolving survey technology into the domain of computer networks have been confronted in various survey procedures. Moreover, an ideal way to gauge what impact a procedure might have in a new medium is to examine its effect on currently used media.

To address the empirical issues concerning the conduct of surveys over a computer network, we examined the literature to see what existing elements might generalize to the domain of network-oriented surveys. Careful attention was paid to what is currently known about surveying via mail and telephone, two media we believe share important features with a computer network. A mail survey (and most likely a computer network survey) is self-administered, relies on a written cover letter to win compliance and confidence, and hinges on well written instructions and carefully crafted questions to guide respondents through the survey. In a telephone survey, respondents cannot see how many questions there are, cannot look ahead (or back) at questions, and cannot easily (if at all) change responses to questions answered, conditions that might also confront respondents to a network survey. We also examined the literature pertaining to computer-administered surveys and any information currently available on surveys conducted over a network.

Our approach, illustrated in Figure 1, was first to extend and adapt what is known to the burgeoning domain of computer network surveys. We perused the literature and identified the knowledge and methodology that is applicable or generalizable to the new media of network surveys. We then identified gaps in the knowledge or procedures, delineated the constraints and advantages of performing surveys on a network, and proposed the most advantageous adaptation for producing active network surveys. We strove to identify the important issues, concepts, and procedures, and also to specify the requirements for developing a survey technology suitable to perform adaptive/automated surveys on computer networks.
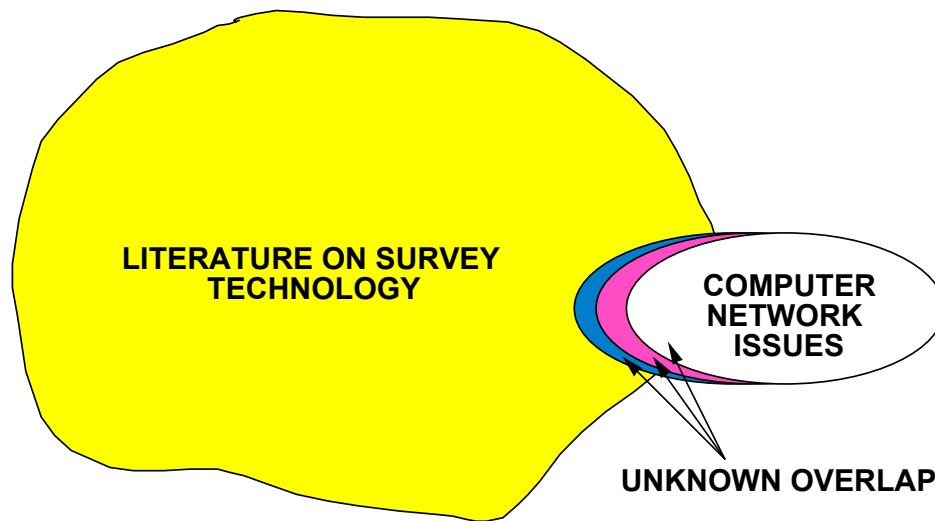
Figure 1. Evaluate Current Survey Literature, Determine Overlap to Computer Network Issues

Document Overview

In this report, we discuss the results of our research effort. The following sections discuss the work carried out to achieve the four objectives we defined for this effort. First, we discuss the evolution of survey technology to conduct in-person interviews, mail surveys, telephone surveys, and the application of computers to survey technology. Second, we present our analysis of survey technology in terms of its application to conducting surveys on computer networks. Next, we delineate what is known from the current literature, what can safely be surmised, and what areas require empirical investigation. Lastly, we describe a pair of pilot experiments designed to investigate presentation issues that emerged from the literature review and analysis. These experiments investigated which of various text and graphics-based format techniques lead to faster, more complete, less error-ridden, and/or more accurate responses. In Appendix A, we present an overview of a research program designed to investigate the issues identified in the experiments.

CURRENT SURVEY TECHNOLOGIES

Traditionally, sample surveys were conducted in person. A surveyor contacted a potential respondent, solicited participation, and administered the survey by reading the questions and recording the responses. This basic process is still very much in use today, but sample survey procedures and methods have not remained static (Buetow, Douglas, Harris, & McCulloch,1996; Johnston & Walton, 1995; Kiesler & Sproull, 1986; Rossi, Wright, & Anderson,1983). Various pressures such as the rising cost of in-person interviews, difficulty in finding respondents at home, and unsafe neighborhoods have led to the development of new survey technologies (Kiesler & Sproull, 1986; Parker, 1992; Sproull, 1986). Two procedures, mail and telephone surveys, have become important tools in the survey industry.

Both procedures address several important issues that arise when considering computer networks as a medium in which to conduct surveys.

Mail Surveys

Mail surveys by their very nature are self-administered. The initiator of the survey must depend on the written word to convince the potential respondent to participate and to respond truthfully and honestly. Written instructions and carefully crafted questions are the only means to guide the respondent through the survey and elicit appropriate responses. These are some of the same conditions that will confront surveys conducted via a computer network.

A nagging problem in the use of mail surveys, particularly when lengthy survey questionnaires are required, has been low response rates (Kiesler & Sproull, 1986; Parker, 1992; Sproull, 1986). Frequently the response rate to a mail survey falls below 15 percent, and the longer the survey instrument the lower the response rate. Surveys implemented by computer networks will probably have to address the same issue, especially with long surveys.

About 25 years ago, survey researchers initiated a serious effort to find solutions to the low response rate problem. One solution to emerge from this research effort was the Total Design Method (TDM) (Dillman, Carlson, and Lassey, 1978), a process that addresses the "look and feel" of the survey. In one review of the TDM's effectiveness, Dillman (1983) reported that with questionnaires averaging ten pages, 28 studies using TDM in its entirety reported an average response rate of 77 percent. In another 22 studies where the TDM was followed to some reasonable degree, an average response rate of 67 percent was attained. Moreover, Dillman reported that no study using TDM had reported a response rate below 60 percent, which is considered very high for a mail survey.

Consistency among all parts of a survey is essential. To focus and apply consistency throughout the survey process, Dillman et al. (1978) followed the tenets of exchange theory (Blau, 1964; Thibaut and Kelly, 1959). The primary assumption is that an individual is most likely to complete a survey when the perceived rewards of participating are maximized, perceived costs are minimized, and the person trusts that the anticipated rewards will be conveyed (Dillman, 1983). General principles followed in constructing TDM designed surveys are from Dillman (1983; p. 362):

-- The questionnaire is designed as a booklet, the normal dimensions being 6.5 x 8.25 inches (16.5 x 21 cm.).

-- The questionnaire is typed on regular sized (8.5 x 11 inches) pages and these are photo-reduced to fit into the booklet, thus providing a less imposing image. Resemblance to advertising brochures is strenuously avoided; thus, the booklets are printed on white paper.

-- Slightly lighter than normal paper (16 versus 20 lb.) is preferred to ensure low mailing costs.

-- No questions are printed on the cover page; it is used for an interest-getting title, a neutral but eye-catching illustration, and any necessary instructions to the respondent.

4

-- Similarly, no questions are allowed on the last page (back cover); it is used to invite additional comments and express appreciation to the respondent.

-- Questions are ordered so that the most interesting and topic-related questions (as explained in the accompanying cover letter) come first, potentially objectionable questions are placed later, and those questions requesting demographic information come last.

-- Special attention is given to the first question; it should apply to everyone, be interesting, and be easy to answer.

-- Transitions are used to guide the respondent from one group of questions to another, much as a face-to-face interviewer would warn of changes in topic to prevent disconcerting surprises.

-- Each page is formulated with great care in accordance with principles such as the following:

- lower-case letters are used for questions and uppercase letters for answers;

- to prevent skipping items, each page is designed so that whenever possible respondents can answer in a straight vertical line instead of moving back and forth across a page;

- overlap of individual questions from one page to the next is avoided, especially on back-to-back pages, with only one question asked at a time in an item;

- visual cues (arrows, indentations, spacing) are used to provide direction.

A few of the issues addressed in the TDM concerning the size and design of a printed questionnaire booklet are not of direct concern when crafting a survey for delivery over a computer network, but the strong underlying issues of presentation design, first impressions, efficiency, and attention to detail are relevant regardless of the particular survey medium (Barron, Tompkins, & Tai, 1996; Comber, 1995; Nielsen, 1996). Dillman cogently argues that survey recipients tend to make holistic evaluations of the survey package, a point that should not be lost when designing network surveys.

Analysis of mail surveys has clearly demonstrated the impact and importance of cover letters--the more individual and personal the better (Carpenter, 1975; Dillman & Frey, 1974)-- and reminders and follow-up contacts (House, Gerber, & McMichael, 1977; Nevin & Ford, 1976; Sproull, 1986). The importance of question and response sequencing has also been shown (Catania, Binson, Canchola, Pollack, Hauck, & Coates, 1996; Krosnick & Alwin, 1987), as well as screen questions or skip patterns in which according to a particular response to a question the respondent is instructed to skip over or answer certain questions.

Telephone Surveys

Another survey technology that shares some similar characteristics with a computer network survey is the telephone survey. The unique aspect of the telephone survey is that respondents are interviewed, as in an in-person survey, but the interviewer is not physically present. Many studies have been conducted (cf. Groves & Kahn, 1979; Herzog & Rodgers,

1988; Herzog, Rodgers, & Kulka, 1983; Jordan, Marcus, & Reeder, 1980) to examine the impact of this situation on the various aspects of the survey process.

Research contrasting face-to-face and telephone surveys has found, for example, that the response rate of face-to-face is higher than telephone surveys (75% versus 69%, respectively) (de Leeuw & van der Zouwen, 1988). De Leeuw and van der Zouwen (1988) also reported findings that speech utterances are longer and the amount of information elicited by open-ended and checklist questions are greater in face-to-face interviews. These differences are most likely due to the superior channel capacity, in terms of both visual nonverbal and audio cues available in face-to-face interviews. But there is also evidence that respondents in telephone interviews may answer more honestly (Krysan, Schuman, Scott, & Beatty, 1994; Sykes & Collins, 1988), their responses are less tainted by social desirability (Cannell, Groves, Magilavy, Mathiowitz, & Miller, 1987; Krysan et al., 1994; Sykes & Collins, 1988), and they are more likely to respond to sensitive questions (Krysan et al., 1994; Sykes & Collins, 1988). Positive results attributable to anonymity and the lower social presence of the interviewer are credited for these latter results and may very well generalize to computer network surveys, as might negative aspects, such as the lower response rate and curtailed responses found in telephone interviews.

Computer-Aided Surveys

Computers have already been widely used throughout the survey domain for data tabulation and analysis. Development of computer-assisted survey technologies was motivated by the desire to make data collection simpler, more efficient, and error-free. Survey researchers praise computer-assisted survey methods because once they are set up, they are easy to use, efficient, and relatively enjoyable and novel, as well as cost-efficient (Anderson & Gansneder, 1995; Buetow et al., 1996; Kiesler & Sproull, 1986; Parker, 1992; Sproull, 1986). For example, computer-assisted personal interviews (CAPI) were developed to facilitate data collection in face-to-face interviews (Buetow et al., 1996). In CAPI the interviewer reads the questions from the screen and types in the responses of the participant. The pace of the interview is set by the computer.

Other ways researchers have begun to employ computers is to assist in administering telephone surveys (computer-assisted telephone interviewing or CATI) and in conducting self-administered surveys (computer-assisted self-interviews or CASI) (Anderson & Magnan, 1995; Buetow et al., 1996; Rodman & Williams, 1996). In a CATI system the interviewer is seated before a computer display screen wearing a telephone headset. The computer presents the questions on the screen in the order they are to be read and in the exact wording to be used. Branching between items is computer controlled and is governed by prior entries or predetermined sequences for a respondent class. Responses are entered directly into the computer by keyboard and can be monitored to detect errors, omissions, or inconsistencies. From the respondents' point of view, it is just a telephone survey, but the pace is probably more consistent since the computer is keeping track of the questions and the interviewer is responsible for reading them. CATI systems provide a means to facilitate or expedite surveys by telephone, making them quicker and easier to complete, as well as the ability to enhance

and control survey data quality (Nicholls, 1988; Groves & Mathiowetz, 1984; Rodman & Williams, 1996).

CATI has made it possible to conduct carefully controlled studies of question wording, question order effects, and response order effects. CATI developers and users have also grappled with how items are to be presented on the display screen (e.g., item-based, screen-based, form-based), methods of questionnaire setup, tailored wording of complex questions based on prior responses, computer-controlled branching between questionnaire items, entry of responses, and automatic range and consistency checking during the survey (Groves & Mathiowetz, 1984; Nicholls, 1988). All these issues will be important and must be addressed when implementing surveys on a computer network. Although in CATI studies the person viewing the screen is the interviewer rather than the respondent, much of what has been found for CATI will generalize to computer network surveys, especially the impact of questionnaire setup, item presentation, and dynamically tailoring the survey according to the interviewee's responses.

The use of CASI, where the individual using the computer to read and respond to the questions is the survey respondent, has grown; and the methodology surrounding its use is maturing. Initially, CASI surveys were similar to self-administered paper-and-pencil surveys, except for the fact they were done on the computer. However, a computer offers many more capabilities that may be utilized to create a more effective survey than paper-and-pencil administrations.

CASI surveys have been enhanced with audio (ACASI) and audio-visual (AV-CASI) cues to guide respondents who may have difficulty with textual presentation. ACASI has allowed surveyors to reach people who are illiterate, or do not have a reading knowledge of the language in which the survey is being conducted (Johnston & Walton, 1995). It is possible to implement ACASI over networks as well, while taking advantage of the adaptive features of network surveys. Similarly, this can be done with AV-CASI. However, the more graphics, animation, and audio features added to a program, the larger the program becomes, and the longer it may take to run or download through a network.

The growth in the use of CASI has also produced opportunities to study its impact on response validity and reliability (Johnston & Walton, 1995; Kiesler & Sproull, 1986; Sproull, 1986; Tourangeau & Smith, 1996). Several studies indicate that computer-assisted self-administered surveys may be associated with more honest responses, less reluctance to answer sensitive questions, and lower socially desirable responses in comparison with personal interviews (Johnston & Walton, 1995; Kiesler & Sproull, 1986; Martin & Nagao, 1989; Tourangeau & Smith, 1996). It is reasonable to expect that surveys conducted over a computer network may enjoy similar positive consequences.

Couper and Burt (1994) report that the respondents' attitudes toward computer-administered surveys are positive. Respondents view them as more scientific, accurate, and secure. However, there is a possibility that individuals completing surveys administered over a computer network may not enjoy the same feelings of confidentiality compared to computer

surveys that do not involve use of networks. This may be due in part to recent publicity about insecure networks and the violations of individuals' privacy because of lack of security. This is an issue that we explore further as we develop guidelines for network surveys.

By taking advantage of the positive aspects of all existing survey technology, it is possible to create surveys that will maximize response rates and data quality as well as minimize errors. The challenge lies in appropriately extending this technology to a new medium—network-administered surveys. In the next section, we analyze the application of existing survey technologies to a new technology.

ANALYSIS OF CURRENT SURVEY TECHNOLOGIES

In this section we describe the results of our functional analysis of current survey technologies. To conduct surveys successfully over a computer network, some existing processes and procedures of surveying may readily be adapted to the new medium while others will require some major changes. The goal of the analysis was to identify what current technologies will carry over with little or no change, what technologies require accommodations that can be derived from the existing literature, and what critical issues must be addressed analytically or empirically to successfully carry out surveys over computer networks. (These latter issues are taken up in detail in Appendix A.) Various constraints imposed by and advantages offered by computers and networks will interact with various aspects of the survey process, and these must be addressed. We identified and focused on these interactions.

Of the surveys currently appearing on the Internet, many are nothing more than mail surveys sent over the network. In these cases the surveyor is exploiting the Internet's free, rapid send and receive capabilities. Surveys conducted as mail surveys over the Internet utilize little of the potential power and capabilities available when surveys are conducted by computer and distributed over a network. The surveys are usually linked to home pages of companies, universities, or organizations such as the American Psychological Society and the American Psychological Association. By attaching a survey to their homepage, the organizations amass a sample and collect data with every 'visitor' to the homepage who chooses to respond to the survey. This is certainly a viable approach, and many surveys will be designed and carried out in this fashion; but it is not a scientific process suitable for conducing a formal survey. If nothing else, this method of self selection is not a suitable method for obtaining a random sample from a population of interest. For more formal surveys, it is likely that surveyors will have the addresses of the intended sample, or at least the general or local address of the intended sample. This could be the general address of, for example, military units, companies, organizations, electronic bulletin boards, and news groups. The survey would be sent directly to the individuals of identified groups.

Moreover, in the surveys that have been conducted on the Internet, little thought has probably been given to the validity and reliability of the data collected in this manner. From past research we know that significant differences exist between the different survey administration modes. Some survey modes do better than others, depending on the domains

surveyed, the approach taken, and the information sought. In our analysis, we used findings of past research to deduce how certain procedures used over a computer network might impact the data collected.

Areas of Low Impact on Transition

We first identify those areas that should make the transition to computer network surveys with little or no alteration. Dillman's (1983) TDM should in large measure be adaptable to surveys conducted over a computer network. Table 1 extends the TDM approach to computer network surveys, showing in column 1 the TDM techniques suggested for mail surveys and in column 2 the adaptation of those rules for the administration of surveys over computer networks.

A number of studies (e.g., Koltringer, 1995; Rodgers, Andrews, & Herzog, 1992; Scherpenzeel & Saris, 1993; 1997) have examined a multitude of different cognitive psychological issues across the various survey administration types in terms of issues of reliability and validity. Many of the findings from these studies should generalize to computer network surveys. It should be noted that measures of reliability and validity discussed throughout this section were usually derived from a multitrait-multimethod design (see for example Scherpenzeel & Saris, 1997). Reliability is akin to Cronbach's concept of internal consistency, and validity refers to correlational estimates of the true score.

Like mail surveys, computer network surveys depend to some significant extent on written introductions and instructions to obtain compliance, convince respondents of the confidentiality of their responses, elicit trust, and impart knowledge on how to perform the survey. We have already noted the importance of the cover letter. Scherpenzeel and Saris (1997) reported that moderate-to-long introductions (i.e., greater than 40 words) as opposed to those shorter in length (i.e., less than 41 words) produced response data higher in reliability and validity. They further report that the optimal arrangement is moderate introduction length paired with longer question length.

Item construction is another ubiquitous task, and for which most (if not all) of the guidelines, methods, and procedures associated with item construction from other survey modes will transition directly to network surveys. Sheatsley (1983) noted that item wording is as much an art as a science, but there are some generally accepted guidelines. Items should ask about one issue at a time, and the use of negatives should be avoided. The effects of question order are still open to research, but it is known that care must be taken when two or more questions deal with aspects of the same issue, or when general summary-type questions are used. Scherpenzeel and Saris (1997) have found that the position of a question in a questionnaire had nonsignificant effects on validity and reliability.

Concerning the type of information asked for, research has shown that questions requesting frequency information have the lowest validity and reliability, agree/disagree statements have low reliability, but high validity, and judgment questions have the greatest reliability (Scherpenzeel & Saris, 1997). These researchers conclude that, in general, type of

information asked for and the balance of the questions have some effect on reliability, but not the validity of the response data.

TABLE 1. ADAPTATION OF DILLMAN'S METHOD FOR MAIL SURVEYS TO SURVEYS ADMINISTERED OVER A COMPUTER NETWORK

Dillman's Method for Mail Survey Administration	Modifications of Dillman's Method for Internet Survey Administration
The questionnaire is designed as a booklet, the normal dimensions being 6.5 X 8.25 inches (16.5 X 21 cm).	The set-up or installation of survey software should be neat and uncluttered. Presentation should not require scrolling across the screen.
The questionnaire is typed on regular sized (8.5 X 11 inches) pages and these are photo reduced to fit into the booklet, thus providing a less imposing image.	Each page fits on the screen, with scrolling kept to a minimum (Comber, 1995; Nielsen, 1996)
Resemblance to advertising brochures is strenuously avoided; thus, the booklets are printed on white paper.	Resemblance to any form of commercialism on the Internet is strenuously avoided. No advertisements should be placed anywhere on the survey or related pages. It should look professional without looking slick.
Slightly lighter than normal paper (16 versus 20 lb.) is preferred to ensure low mailing costs.	Not applicable
No questions are printed on the first page (cover page); it is used for an interest-getting title, a neutral but eye-catching illustration, and any necessary instructions to the respondent.	The initial page on an electronic survey should contain an interest-getting title, a neutral but eye-catching illustration and/or logo, any necessary instructions to the respondent, and a link to begin the survey.
No questions are allowed on the last page (back cover); it is used to invite additional comments and express appreciation to the respondent.	No questions are allowed on the last screen that is displayed; it is used to thank respondents for participation, and contains a form respondents can use to send comments to the experimenter if they wish.
Questions are ordered so that the most interesting and topic-related questions (as explained in the accompanying cover letter) come first; potentially objectionable questions are placed later, and those requesting demographics information last.	Questions are ordered so that the most interesting and topic-related questions (as explained in the introductory web page) come first; potentially objectionable questions are placed later, and those requesting demographics information last.
Special attention is given to the first question; it should apply to everyone, be interesting, and be easy to answer.	Directly applicable
Transitions are used to guide the respondent from one group of questions to another, much as a face-to-face interviewer would warn of changes in topic to prevent disconcerting surprises.	Directly applicable
Only one piece of information is asked for per item	Directly applicable
Lowercase letters are used for questions stems and uppercase letters for response options.	Directly applicable
To prevent skipping items, each page is designed so that whenever possible respondents can answer in a straight vertical line instead of moving back and forth across the page	Provide a link on every page to move forward and, if appropriate, a link to move backward.
Avoid overlap of individual questions from one page to the next, especially on back-to-back pages	Do not overlap individual questions from one screen to another.
Visual cues (arrows, indentation, spacing) are used to provide direction.	Visual cues and animated graphics can be used to provide direction.

In terms of response categories and response scales, Scherpenzeel and Saris (1997) found that the symmetry of the response scale had nonsignificant effects on validity and reliability. Only the existence of an explicit midpoint was shown to have a moderate effect on validity, but here the conclusion is simple: use an explicit midpoint whatever the survey mode. Scherpenzeel and Saris (1997) also found that the direction of the first presented category of a scale has some effect on reliability, but not the validity of the response data. A frequent concern in questionnaire construction is whether a "don't know" or a "not

applicable" response category should be concluded. Scherpenzeel and Saris (1997) reported that whether such a category is included has little impact on either validity or reliability of the survey instrument.

In some cases, the impact of question type on reliability may be changed when the survey is administered on a computer. For example, the length of the response scale is often cited as one of the most important survey features contributing to validity and reliability. Observations of computer-administered surveys (e.g., CASI, CAPI) show that when respondents must choose and type in the number connected with a response category as opposed to sliding a cursor across the screen and marking the selected option, more mistakes are made and reliability suffers (Saris, 1991). Computer network surveys will do well to use slider bar or similar scales and limit the range of the scales from 0 to 10, a procedure linked to high validity and reliability as demonstrated by the pilot studies conducted under this effort and in previous research (Scherpenzeel & Saris, 1993).

As we indicated above, the use of computers has impacted every survey administration mode. The computer-assisted personal interview (CAPI) has become the most commonly used method of face-to-face data collection (Tourangeau & Smith, 1996). The use of computer-assisted self interviews (CASI) is growing rapidly in popularity as the way to obtain responses to sensitive issues (Couper & Rowe, 1996). Because any survey delivered over a network is essentially conducted by computer, CAPI, CASI, and other computer-assisted survey methods can provide some insight into the issues involved in conducting surveys over a computer network.

A number of studies have evaluated or compared various computer-assisted survey procedures (cf. Couper & Rowe, 1996; Tourangeau & Smith, 1996). Some of the virtues touted for CAPI are the same for network surveys: improved data quality, faster delivery, and lower cost. But research findings caution that the mere use of computers does not guarantee data quality. Data quality is bound to suffer if individuals completing computer-aided surveys are not experienced with computers, are impaired in some way inhibiting easy use of a computer (e.g., vision problems making it difficult to see the display, arthritis making it difficult to control the mouse), or are not literate. Couper and Rowe (1996) have noted that the number of minorities (particularly non-white respondents) is often positively correlated with lack of computer experience and literacy. As with CATI, CAPI, and CASI, the capabilities of computer-conducted network surveys to adapt to individual proclivities must be marshaled to cope with such problems.

On the positive side, despite technical difficulties that can arise when computers are employed, the attitude of respondents toward newer computer-assisted survey technologies is positive (Couper & Burt, 1994). It is also reported that respondents tend to consider computer-administered surveys more scientific, more accurate, and more secure.

Areas of Higher Impact on Transition

One aspect of the TDM that cannot be assumed to generalize from printed surveys to computer surveys is the issue of how items should be transmitted to the respondent's computer system and formatted on the display. An important issue (identified for future research in Appendix A) that will require systematic research is whether surveys should be presented in part (one or a few items at a time to the screen in a serial fashion) or whole (the whole survey to the screen). Table 2 shows the advantages and disadvantages of part and whole presentation of surveys as well as guidelines for screen layout in each mode. Note that we are assuming a mail-type survey purveyed over a computer network and simple net software on the recipient's side to receive it. The relative importance and consequences of these positive and negative factors must be empirically determined.

TABLE 2. COMPARISONS OF PART AND WHOLE PRESENTATION OF SURVEYS

	Part Presentation	Whole Presentation
Advantages	• Supports error-free branching • Keeps the screen uncluttered • Minimizes scrolling • If respondent quits midway, data to that point is recoverable	• Very easy to implement • Once downloaded, the whole survey is local to respondent's machine
Disadvantages	• May be slow, especially over networks • Cannot go backwards easily if mistake is made	• Hard to make survey adaptable; must send with Java or similar applet • Lots of scrolling may be required, especially if survey is long • If respondent quits midway, all data are lost
Appropriate if	• Survey contains a large number of branches • It is not necessary or appropriate to move backward in the survey	• Branching in the survey is difficult without applet (usually, all items are to be answered by all respondents) • The survey is relatively short (roughly less than 3 screens) • Completion will be helped by respondents' knowing survey length • It is advantageous for respondents to move forward and backward while answering the survey
Layout Guidelines		• Small groups of related items or all items in a scrolling window • Little or no branching

Electronic surveys have yielded greater completion rates and fewer item-completion mistakes compared to their pencil-and-paper counterpart (Kiesler & Sproull, 1986). However, this result may accrue because at present electronic surveys are usually completed by a self-selected sample of individuals who have access to computer networks and are highly computer-literate. Here the issue of response accuracy for network-administered surveys is open. Careful formatting and creative use of help should support high completion and low error rates.

Table 3 presents a proposed set of guidelines that are derived from our analyses addressing issues relevant to conducting effective surveys over the Internet. The guidelines concern the design of the response format, the method of transmitting the survey over a network, and adaptation mechanisms that will support the respondent in completing the survey effectively, efficiently, and accurately. They address areas of concern that arise when computer surveys are delivered to and completed by individuals who are not highly sophisticated computer users, and who may, in some cases, be unenthusiastic about, and perhaps even fearful of, using a computer.

TABLE 3. PRESENTATION AND ADMINISTRATION GUIDELINES:
EXTRAPOLATION TO NETWORK SURVEYS

Response Format:

– Use graphics, labor-saving aids, and encouragement.
– Keep response scale length to under 12 when using slider bars or similar response

mechanisms.

- For multiple choice items, respondents should be able to check off appropriate number of responses. Provide error checking upon confirmation (e.g., too many responses are checked off, no answers are checked off, etc.).
- Provide a text box for open-ended questions

Network:

- Make sure the survey and related software are easy to install.
- Make survey compatible with as many platforms as possible. (Note: using Java precludes using text-based web browsers and older versions of Netscape and Mosaic.)
- If there is a possibility respondents might not have the necessary software or plug-ins to do the survey, create a link for them to download and install it (make sure copyright laws are addressed).
- Assure net responders confidentiality: Use appropriate networking hardware and software to provide what confidentiality is possible.

Adaptation Mechanism:

Help or explanation buttons:

- For definition of key terms.
- For simpler or alternative wording or audio version of item.
- To explain response scale.
- Reiteration of instructions.
- Other applicable information

Animation, Color, and Graphics:

- Use to attract and/or hold attention.
- Use illustrations or examples to clarify what is asked/wanted.
- Prevent overuse so animation, color graphics, etc. are not distracting.

In terms of adaptation mechanisms, we note that there is a paucity of studies evaluating the impact of formatting issues, graphics, animation, and aids/help functions on response data reliability and veracity. One of the true virtues of computer-administered surveys is the ability to implement all the aforementioned procedures, yet little empirical evidence exists on how best to employ them and what effect they will have on the response data. The use of appropriate aids, help options, and encouragements can work to improve the reliability of a survey instrument. Appropriate graphical aid such as those tested in the pilot experiments (see following section) can reduce such mistakes and improve reliability.

In exploring such issues, the domain of human-computer interaction offers many potentially applicable guidelines issues which are germane to computer-administered surveys. Extensive research has been done in the past two decades on issues surrounding the display of

information on a computer screen and on devices and methodologies for inputting information into a computer. Researchers in this domain have also investigated the use of "help" panels and menus. Some of the guidelines which are applicable to computer surveys are presented in Table 4 (from Helander, 1988).

TABLE 4. GUIDELINES FROM HUMAN-COMPUTER INTERACTION ON SCREEN
DESIGN ISSUES AND TECHNIQUES

Aspect of Presentation	Related Guidelines
Amount of Information to Present	Make appropriate use of abbreviations.
	Avoid unnecessary detail.

	Use concise wording. Use familiar data formats. Use tabular formats with column headings.
Grouping of Information	
Color	Presenting different sets of display elements in contrasting colors clearly creates some degree of grouping within elements of the same color. Proximity of elements will make the visual association stronger.
Graphical Boundaries	A common technique for grouping items is drawing graphical boundaries around related elements.
Highlighting	Another way of creating visual groups is the use of highlighting, increased brightness, or reverse video for related elements.
Highlighting of Information	
Reverse-video	This can be used to highlight a group of elements to draw attention to a particular portion of the screen
Color	Presenting a screen element in a different color from the rest of the elements attracts attention.
Underlining	Underlining words within a large block of text draws attention to those words.
Flashing	Flashing can draw attention to a screen element, but causes annoyance to users if it cannot be turned off.
Placement and Sequence of Information	
Sequence of Use	If items must be responded to in a certain sequence, they should be presented in that order.
Importance	Present absolutely crucial items for users to respond to early in the sequence.
Generality/Specificity	More general items should precede the more specific items in a section.
Spatial Relationships among Elements	
Indentation	Subordinate or hierarchical relationships among items can be conveyed effectively through the use of indentation.
Process Associations	Using computer graphical displays to represent actual elements of a process makes the task and its status more clear to the user.
Presentation of Text	
Letter Case	Traditional mixed upper and lower case is easiest to read. All uppercase is used to highlight key words.
Justification and spacing	Allow ragged right margins instead of "fill justification."
Spacing between paragraphs/sections	Leaving blank lines between items facilitates readability.
Line Length	Lines should fit on the screen so left-right scrolling is not necessary.
Uses of Graphics	
Representing Numerical Data	Representing numerical data pictorially makes it easier to read and understand (e.g., pie charts, simulated measures).
Representing Direct-Manipulation Objects and Actions	Use of icons to represent real-world objects makes learning an interface more intuitive to the user.

PILOT EXPERIMENTS: METHODOLOGY AND RESULTS

Experiment 1

Purpose.

The purpose of Pilot Experiment 1 was to investigate response format effects in Computer–Aided Self Interviewing (CASI) surveys such as would occur within establishment

surveys administered via a computer network. The manipulations in this study involved providing survey respondents different degrees of assistance and encouragement while they were answering a series of questions posed in formats that are highly attention–demanding and somewhat lengthy. We supplied half the respondents with "aids" and "enhancements" to ease the burden, hopefully leading to faster, more complete, less error-ridden, and/or more accurate responses (we cannot, however, determine the actual veracity of the responses with this design). The goal was to determine whether the assistance and encouragement provided were helpful in stimulating respondents to continue giving careful as opposed to lackadaisical answers and whether any such effect persists through a series of questions. We also examined how format affected the four response formats (termed "tasks") that were administered once in Block 1 and then repeated in Block 2.

Subjects

Participants for this experiment were 41 male and female undergraduate students who were remunerated for their participation. Subjects were randomly assigned to one of two Questionnaire Format groups : Ordinary (n = 21) and Enhanced (n = 20). The procedures of the study were reviewed and approved by the campus Institutional Review Board (IRB).

Independent Variables.

Questionnaire format (Quexform) The design for Pilot Experiment 1 manipulated Questionnaire Format over two between-subjects conditions: ordinary and enhanced. The **Ordinary** format was designed to resemble the static quality of paper-and-pencil questionnaire administration. As such, an item was placed on the computer screen and a blinking cursor showed where the typed response was to appear. When the respondents finished answering, they answered a probe that they were ready to move to the next question. The computer program did nothing to ease the respondents' workload and did minimal error checking before accepting an answer.

The **Enhanced** format was designed to be more dynamic in its interaction with respondents, resembling a growing proportion of today's computer software. Various supports or aids were provided to respondents for each task to ease their workload while answering the questions. For example, in the percentage allocation task (Pct), a running tally of points already allocated was provided, so that respondents did not have to do sums in their heads. This format also provided words of encouragement.

Respondents were randomly assigned to either the Ordinary or Enhanced condition and were unaware that another form of the questionnaire even existed. The null hypothesis for the Ordinary vs. Enhanced contrast is that there is no difference in the means of the dependent variables between conditions; the alternative hypothesis of scientific interest is that respondents in the Enhanced format condition will give "better" answers (more complete, fewer errors, faster) than respondents in the Ordinary format condition.

Block of questions. The eight questions in each task format were divided into two fixed sets of four. One set of four from each task format was placed in the Block 1, making a total of 16 questions (four questions per task times four task formats; see Fig. 4). The remaining sets of four not yet used were placed in Block 2, again yielding a total of 16 questions. The Blocks were then counterbalanced for order. Half the respondents did Block 1 and then did Block 2. The order of the two Blocks was reversed for the remaining half of the respondents (Block 2 followed by Block 1). The order of questions within a given set was randomized.

Blocks is a within-subjects (or repeated measures) factor, as respondents experience both conditions, that is, they answer the questions in Block 1 *and* the questions in Block 2. The null hypothesis for the Block factor is that there will be no difference in the means of the two Blocks for a specific question format; the alternative hypothesis of scientific interest is that there is a difference in performance between Blocks. For example, a downward change (poorer performance) might be caused by fatigue, whereas an upward change (improved performance) might be caused by practice.

Dependent Variables.

Task Formats. We developed four Task formats so that we could study a variety of the ways in which questions are typically posed in paper-and-pencil and computer-assisted surveys. These formats were: (1) generate a list of items within a category, such as listing as many types of fuel as possible ("List"); (2) ranking a series of items from most to least preferred ("Rank"), such as ranking components of job performance (examples of the text-based ranking task under the Ordinary and Enhanced Format conditions can be found in Figs. 2 and 3, respectively); (3) endorsing or checking off all the items in a list that are judged applicable ("YN"), such as whether the respondent uses specific coping mechanism to deal with personal problems; and (4) allocating a total of 100 percent to a series of categories, such as how much of a state's budget should be allocated to different activities ("Pct"). The List task is free response in the sense that respondents must produce the list of items, whereas Rank, YN, and Pct call for respondents to work with items provided in the survey instrument. The Rank and Pct tasks use numerical responses and respondents must keep track of the ranks used or the total percent allocated, respectively. The YN questions are presumably the least burdensome, as respondents do not need to produce alternatives or keep track of prior responses–only give a simple Yes or No answer. Thus, this set of four question formats provides a wide sampling of levels and types of respondent workload in question answering. Eight items in each question format were developed.

Suppose you were being transferred to a new location. Rank order the importance to you of the location characteristics listed below. Under the 1, type the number of the item that is most important. Under the 2, type the item that is next most important, and so on.

    1 Climate
    2 Closeness to mountains
    3 Closeness to water (ocean, lakes, river)
    4 Distance from important relatives or friends

```
    5   Dominant political climate
    6   Ethnic diversity
    7   Housing prices
    8   Quality of schools
    9   Local economy
    10  Presence of airport
    11  Presence of related industry
    12  Region of the country
    13  Region of the world
    14  Size of the town/city


 1  2  3  4  5  6  7  8  9  10  11  12  13  14
 4  5  3  6  7  1  2  8  9  10  13  14  11  12


Press N key to revise your answer
or press Y key for next question.
```

Figure 2. Example of Text-Based Ranking Task Screen Under the Ordinary Format

```
THIS GROUP OF QUESTIONS INVOLVES ASSIGNING A RANK NUMBER TO A LIST
  OF ITEMS WHICH WE PROVIDE.

  IT IS IMPORTANT TO REVIEW THE LIST TO BE SURE YOU HAVE ASSIGNED A RANK
  TO EACH ITEM. ALSO CHECK TO SEE THAT YOU HAVE NOT ASSIGNED THE SAME
  ITEM TO MORE THAN ONE RANK.

  WHEN YOU ARE READY, PRESS ANY KEY TO CONTINUE...
```

Suppose you were being transferred to a new location. Rank order the
  importance to you of the location characteristics listed below. Under
  the 1, type the number of the item that is most important. Under the
  2, type the item that is next most important, and so on.

          1 Climate
          2 Closeness to mountains
          3 Closeness to water (ocean, lakes, river)
          4 Distance from important relatives or friends
          5 Dominant political climate
          6 Ethnic diversity
          7 Housing prices
          8 Quality of schools
          9 Local economy
         10 Presence of airport
         11 Presence of related industry
         12 Region of the country
         13 Region of the world
         14 Size of the town/city

      1 2 3 4 5 6 7 8 9 10 11 12 13 14
      1 2 3 4 5 6 8 10 11 10 9 7 13 14

Please carefully compare your list to the original list of items.
  Be sure you have assigned a rank to each item in the original
  list and you have NOT assigned an item more than one rank. Press
  KEY Y to revise your list; press KEY N for the next question.

Figure 3. Example of Text-Based Ranking Task Screen Under the Enhanced Format

Primary responses. For each Task format a particular feature of the answers collected
was tallied. The idea was to measure the quality of the respondents' answers. For List
questions we tallied the number of non-redundant responses generated (effort). For Rank
questions we tallied omitted or duplicate ranks (i.e., errors). For YN questions we tallied the
numbers of items checked (level of endorsement). For percentage allocation questions we
tallied the number of points allocated above or below 100 (errors).

Response times. The time expended to complete each Task format was recorded in
seconds and was analyzed in the same manner as the primary responses.

Mood. Fifteen questions designed to assess mood were administered to respondents
after Block 1 and again after Block 2. The subjects made their responses on 11-point bipolar
rating scales. The full 15-item mood scale and subscales developed subsequently were used
in the analyses.

Background variables. Ten background variables were collected from respondents at
the end of the study. Of particular note was the item assessing frequency of computer or
keyboard use. Analysis showed that the Quexform conditions differed as to this item: the
subjects in the Enhanced condition reported less familiarity with computers than those in the
Ordinary condition. Analyses using this item as a covariate were conducted. Although in

several analyses the regression covariance was significant, at no time did the covariate alter the pattern of results revealed when a covariate was not used. For this reason, no covariate results are presented.

Design and Procedure

As Figure 4 shows, Questionnaire Format is a between–subjects factor, whereas Blocks and Tasks are within–subjects factors. For each task format there were four questions or items. The questions concerning motivation and mood were asked after each Block.

					BLOCKS			
QUESTIONNAIRE FORMAT				1			2	
			TASKS				TASKS	
	A	B	C	D	A	B	C	D
1—**Ordinary**: no encouragement, feedback, prompts, or guides		[4 items for each task format]				[4 items for each task format]		
2—**Enhanced**: encouragement, feedback, prompts, and guides		[4 items for each task format]				[4 items for each task format]		

Figure 4. Experiment 1 Design.

The experiment was administered using *Ci3* (Sawtooth software of Institute for Social Science Research) survey software system. This software offers a graphical interface, but in this experiment we presented content primarily in a textual manner. The software provides ready–to–analyze data files at the conclusion of data collection. Each respondent sat at a computer workstation and completed the items of the survey appearing on the screen, using a keyboard and mouse.

We employed additive scales throughout; that is, scales were formed by summing relevant items. In the case of the primary and time responses, the scales were predefined, so there was no question as to which items belonged where. For example, the four List items in Block 1 (which were different for various groups of respondents) were summed up, as were the four items in Block 2. These two variables, and corresponding variables for the other task formats, were submitted to analysis of variance. The reliabilities (Cronbach alphas) for these dependent variables ranged from .70 to .90, indicating that the reliability of these dependent variables is satisfactory.

Mood items received special treatment. Through a series of factor analyses, three subscales were developed from the mood items; these were dubbed *tiresomeness* of the task (4

items: "Very Boring," "Should End," "Not Very Likable," and "Should Not Continue"), *difficulty* (4 items: "Not Very Comfortable," Very Difficult," Not Very Easy," and "Need More Instructions"), and *pace* (3 items: "Very Slow," "Not Very Fast," and "Need Fewer Instructions"). The scales are almost uncorrelated and can therefore be considered separately. The reliabilities of the 15-item full scale and of the subscales were satisfactory, ranging from .64 to .84.

Results

Analyses of mood scales. Analysis of Variance (ANOVA) revealed a Block effect for the full 15-item mood scale ($F(1,39=13.28$, $p<.001$); Quexform and Quexform x Block were not significant. When the three subscales were examined using ANOVA, a Block effect was found for tiresomeness ($F(1,39)=8.75$, $p<.01$); no other effects were significant for tiresomeness, difficulty, or pace. These analyses were repeated using nonparametric procedures, in case normality assumptions of the parametric test were not fully met. Friedman tests confirmed the Block effects for the 15-item full scale ($p<.005$), tiresomeness ($p<.05$), and difficulty ($p<.05$), and showed a trend for pace ($p<.10$).

Since lower numbers are more negative and means declined from 6.46 in Block 1 to 5.93 in Block 2, respondents' "moods" deteriorated somewhat over the course of the study. Respondents found Block 2 more tiresome than Block 1, and the nonparametric tests showed similar results for difficulty and pace. These results appear to reflect fatigue and possibly irritation, which was as expected given the attention to long items that the study required.

Analyses of the dependent measures A second assumption of the ANOVA is normally distributed errors. Data collected from human and animal subjects often contain a few observations that appear quite different from the main body of observations, termed "outliers." Sometimes a subject misunderstands instructions, or falls asleep at the switch, or gives intentionally faulty data for one reason or another. As these outliers can greatly affect results, standard practice now invites examination of results following their removal. Where outliers made a difference, it is noted in the analyses below.

Error. The error measure for the allocation (Pct) task exhibited a strong Quexform effect as shown in Figure 5. (F(1,39)=15.24, p<.001). Respondents made a number of errors in the Ordinary format condition, but almost none at all in the Enhanced condition. Moreover, this effect was unchanged by removal of two outliers (F(1,37)=15.11, p<.001). Mann-Whitney nonparametric tests contrasting Quexform conditions within Blocks 1 and 2 confirmed this *effect (*p<*.0001 and* p<*.001, respectively).* There were no Block or Quexform x Block effects.
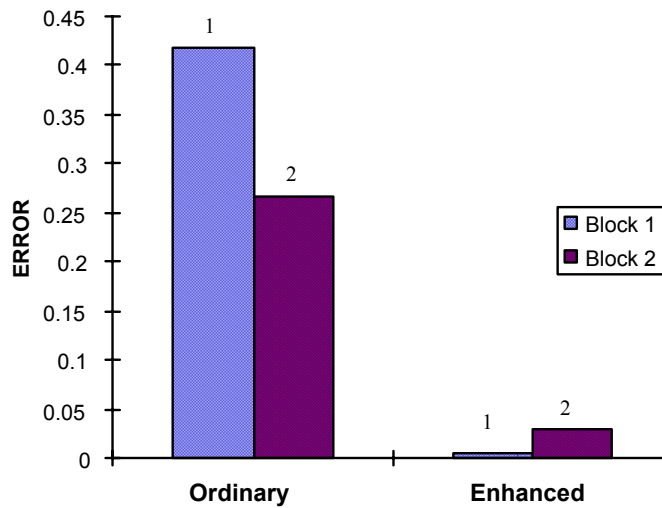
Figure 5. Percent Error Expressed in Arcsin for Ordinary and Enhanced Questionnaire Formats

The Rank error dependent variable exhibited no main effects for Quexform or Block, but did show a Quexform x Block interaction (F(1,39)=6.02, *p<*.05). Inspection of Figure 6 and the means shows that the Enhanced format produced fewer errors than the Ordinary format in Block 1, whereas in Block 2 the means for the two conditions were about the same.
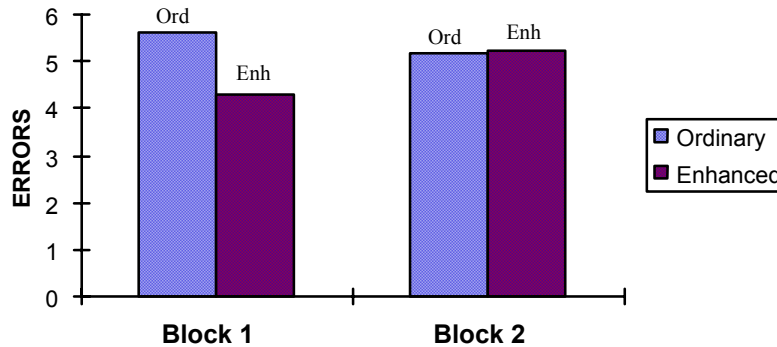
Figure 6. Count of Errors Using Freeman-Tukey Transformation for Ordinary and Enhanced Questionnaire Formats Within Blocks

Analysis of the errors associated with the YN task exhibited a marginally significant effect for Quexform (F(1,39)=2.85, *p<*.10). A Mann-Whitney test showed a marginally significant effect (*p<*.10) indicating lower errors for the Enhanced condition for Block 1 only. There were no Block or Quexform x Block effects.

The MANOVA analysis of the List dependent variable revealed no main effects or interactions and parallel nonparametric tests yielded similar results. The variable computer familiarity (Background question 4) produced a very strong covariate within-cell regression (p < .005), indicating that those who use computers less frequently also produced shorter lists of items. Adjusted for the covariate, the Quexform factor was marginally significant (F(1,38)=2.96, p<.10). The adjusted mean for the enhanced condition was 7.5 percent larger than the adjusted mean for the ordinary condition.

Response time. Results from the Allocation, Rank, and List tasks all showed the same pattern of results: a marked drop in completion time from Block 1 to Block 2 (F(1,39)=16.61, p<.001; F(1,39)=6.41, p<.05; F(1,39)=22.53, p<.001; respectively).[1] In each case the effect was somewhat larger still when outliers were removed, and in each case a Friedman test further confirms the Block effect. It appears that participants were able to perform the Allocation, Rank, and List tasks more quickly the second time around. There were no Quexform or Quexform x Block effects for these three tasks

Discussion.

The primary hypothesis was supported. Respondents produced better quality responses using the Enhanced Format. Clearly the aids and encouragements of the Enhanced Format reduced the errors participants made performing a variety of somewhat labor intensive task when compared to the Ordinary Format. The aids and encouragements used for the Enhanced Format were all text-based and rather restrained. We believe that if we took full advantage of the options afforded by CASI, such as graphics, color, and/or animation, we would have achieved even stronger results for the Enhanced Format**.**

Although the mood measures showed that participants were not as enthusiastic performing the second Block of tasks as compared to the first Block, there was little evidence that the decline in positive mood negatively affected their performance on the second Block. Participants tended to finish the second Block faster, and there was no indication that they made more errors performing the second Block. The Enhanced Format was so efficient in reducing errors (in both Blocks) that there was not any room for improvement, thus we could not assess if the Enhanced Format combated the effects of negative mood.

If we assume that the Ordinary Format is a surrogate for pencil and paper formats and the Enhanced Format represents the added power of computer-mediated surveys (as would

---

[1] Response time for Yes/No items was not available for analysis owing to an error programming the Ci3 survey CATI system.

also be true for computer networks), then the benefits of format enhancements are clear. At a minimum we can achieve better--more error-free and faster--responses using CASI surveys performed via a computer network. Moreover, we have only begun to scratch the surface of such opportunities and benefits.

<u>Experiment 2</u>

<u>Purpose</u>

The goal of Pilot Experiment 2 was to study the impact of graphical user interfaces (GUIs) on the performance of relatively difficult survey items and to pilot methodology that might prove suitable for the conduct of further studies concerning specific design issues that arise in planning and implementing Internet-based surveys. Experiment 2 builds on several features of Experiment 1, which demonstrated that a computer-administered text-based survey equipped with specific aids and enhancements produced significantly more reliable and error-free response data than surveys not so equipped.

As in Study 1, we selected survey items that impose a rather large memory or computational burden on respondents, but we presented these items using a Guided User Interface (GUI) with "enhancements" that were intended to ease some of the burden. For example, in the rank-ordering format, movable tiles carrying the ranks 1 through 15 were displayed. The respondent could use the mouse to move a rank-tile next to each item in the list to be ranked ("drag and drop"). Under this method no rank or item can be omitted (the interface will not advance to the next question), no rank can be assigned twice, and the user can readily see what has been done and what is left to do. Moreover, it is easy to visually inspect one's answers and make final adjustments to the rankings–even after all 15 tiles have been moved, a respondent can still easily make changes by moving the tiles with the mouse.

The ultimate test of a survey methodology is its ability to produce accurate information, but accuracy can only be examined empirically in the unusual circumstance that the surveyor already knows the "true" answers to the questions posed. However, an essential ingredient of accuracy is test-retest reliability, the fact that the same answer is given to the same question on two occasions. When reliability is absent, there can be no accuracy, because two different responses have been received to a question with one true answer. (One danger with an interactive interface can be that respondents treat it as a computer game, not considering their answers but instead trying to "play fast.") In Experiment 2, respondents faced the same questions twice at a two-day interval so that we could investigate test-retest reliability of responses collected with our new GUI question formats.

The final element in Experiment 2 was the use of respondents currently serving in the Reserves or the National Guard, most in enlisted status, many of whom had completed a term of active duty in the regular forces prior to joining the Reserves. By varying the populations from which our samples were selected, we endeavored to demonstrate the robustness of our procedures and their applicability to military subjects.

<u>Subjects</u>

Respondents were recruited through advertisements in the campus newspaper and through flyers posted on campus bulletin boards. A total of 19 current members of the Reserves or National Guard volunteered to be included in the study. A total of 14 completed Trial 1, and 12 of those completed Trial 2. Respondents were paid $10 per hour for their participation. Payment was made at the completion of Trial 2. Prior to beginning recruitment, the Experiment procedures were reviewed and approved by the campus IRB.

<u>Independent Variables</u>

<u>Item format</u>. Three of the task formats employed in Experiment 1 were modified for use in Experiment 2. The Yes-No format illustrated in Fig. 7 was implemented graphically and manipulated over two item formats: in the **YN** version respondents are required to "check all items that apply" as yes's and all those that do not apply as no's; whereas in the **Y0** (Y-null) version all those items not checked as yes's are assumed to be intended as no's. The YN version presumably induces greater control and care at the cost of some speed, as compared with the less labor intensive Y0 version.

A second pair of item formats was derived for a ranking task: in the **Rank-Left** (**RL**) version depicted in Fig. 8 tiles carrying ranks are moved by mouse in a "drag and drop" fashion from the right side of the screen leftward next to the to-be-ranked items; in the **Rank-Right** (**RR**) version shown in Fig. 9 rectangular fields carrying the items' texts were moved rightward into alignment with rank-tiles. That is, either rank-tiles were moved leftward (RL), or they were moved rightward to the ranks (RR). Cognitively it may be more natural to move the ranks to the items to-be-ranked as in the RL version than to move the item's text to the ranks as in the RR version.

Figure 7. Example of the YN Version

Figure 8. Example of the Rank Left Version

Figure 9. Example of the Rank Right Version

Figure 10. Example of the Allocate Version

28

For an allocation task, the item formats were manipulated over two versions of allocation: percentage points or fixed quantity. In the **Allocate Percent (AP)** version shown in Fig. 10, the quantity was 100 percentage points, whereas in the **Allocate Money or Time (AM)** version the fixed quantity was a total dollar amount to be disbursed for various items, or a total number of hours (e.g., 40 in a work week) to be spent on various activities. It was unclear if respondents would find it less cognitively demanding to allocate percents or a fixed quantity of something.

Finally, slider bars (seeming horizontal rods with a slip-ring marker that can be moved along the rod with the mouse in drag-and-drop fashion) with differing range lengths were compared. Figure 11 illustrates the short version or 0-10 length scale that was contrasted with a long version or 0-100 length scale to assess respondents' satisfaction or frustration (the "mood" questions from Experiment 1) with the three tasks formats. Research has indicated that response scales 0-10 or less in length tend to be more reliable with no detriment to validity than response scales that are much longer, e.g., 0-100 in length (Scherpenzeel & Saris, 1993).

Trial. The same questions were asked of respondents on two occasions, or trials, two days apart (e.g., on Monday and Wednesday). Trial is thus a within-subjects factor for those analyses that compare the levels (means) of responses on the two separate days. It is also the interval over which reliability was assessed.

Figure 11. Example of the 0-10 Slider Bar

Blocks of questions. As in Experiment 1, the 8 questions in each task format (Yes-No, Ranking, Allocation) were divided into two sets of four questions each which corresponded to the two versions for each format (YN and Y0, RR and RL, and AP and AM). A block of questions consisted of three sets of four questions from the three task formats, arranged in random order within set and between sets, followed by the 15 mood questions that employed one of the two versions of the slider bar. (Further randomization would have been possible, such as mixing the sets between Blocks, but this would have undermined the examination of the reliability of the mood questions, which were intended to assess mood after a fixed prelude of 12 earlier questions–if the prelude were changed, a lower reliability might have been due to changed conditions rather than a measurement problem.) The two kinds of Blocks thus consisted of:

| Block 1 | [1 2 3] | Mood XY |
| Block 2 | [4 5 6] | Mood YX |

where brackets indicate randomized ordering on each trial. The order of the Blocks was counterbalanced with order 1-2 at Trial 1 and order 2-1 at Trial 2 for 7 respondents, the reverse for 5 respondents (the two "no-show" respondents at Trial 2 were in the latter group). Our use of randomization and counterbalancing in the experimental design means that order effects are eliminated as an explanation of resulting aggregate patterns in the data.

Dependent Variables

Primary responses. The responses collected from each respondent on two occasions were correlated in order to assess test-retest reliability. In the case of YN and Y0 item formats, the response to a single question was a series of 15 1's (Yes's) and 0's (No's) corresponding to the 15 items that might have been checked. We used a form of correlation appropriate to a 2x2 contingency table that characterizes the "agreement" between the two occasions. That is, when an item was marked either as a 1 on both occasions or as a 0 on both occasions, there was agreement, which amounts to reliability. For the rank-ordering item formats RR and RL, a rank order correlation of the rankings for the two occasions was used to assess reliability. Rank order and Pearson product-moment correlations were used to assess the reliability of the Allocation questions, AP and AM, as well as for the 0-10 and 0-100 scales of the mood questions.

For each task format, a particular feature of the answers was tallied. Unlike Experiment 1, these features could not be response errors (e.g., omitting or double-using a rank, or allocating percents that do not add to 100) because our GUI item interfaces all but did away with errors by checking for errors and by requiring they be corrected before advancing to the next question. Thus for YN and Y0 the percentage of items checked "yes;" was tallied for RR and RL, the extent to which the rank ordering corresponded to the initial serial order of items; and, similarly, for AP and AM, the extent to which earlier items were allocated greater shares of percentage of money or time wastallied.

Response times. The time taken to respond to each question was recorded in seconds, transformed to natural logs as in Experiment 1, and analyzed. We expected a decline in response times from Trial 1 to Trial 2, as observed in Experiment 1. We also expected to obtain an estimate of relative response times for different versions of the same GUI item format.[2]

Mood. The 15 "mood" questions were administered to respondents at the end of Block 1 and again at the end of Block 2. These were bipolar rating scales with a response range scale length of 0 to 10 in Block 1 and a response range scale length of 0 to 100 in Block 2. They were intended to assess respondents' affective reactions to the previous 12 questions in the Block. As such they offer another means of assessing the GUI question formats, since respondents who find the formats enjoyable are likely to continue to the end of a questionnaire and try to give thoughtful answers, whereas respondents who become frustrated are more likely to quit or to put less effort into answering the questions. Thus, the mood questions were an adjunct to our other ways of assessing whether the GUI formats are suitable for use in computer network-based surveys.

Background variables. A range of background variables was collected from respondents at the end of the study. Based on our experience from Experiment 1, we included somewhat more detailed questions on the frequency with which respondents worked with computers and performed arithmetic calculation. Other questions addressed attention to detail, concern with the task as a whole, and organizational skills. We also obtained educational background, previous military experience, and current military job in the Reserves or National Guard.

Design and Procedure.

The experimental design was wholly within-subjects. That is, each respondent experienced all tasks, all item versions, and Blocks on each of two occasions (trials). Similar to Experiment 1, Experiment 2 was administered using the *Ci3* survey software system, employing its graphical interface capability. Most respondents completed each trial in just under an hour, though a few took a bit longer on the first day. All 12 participants agreed to be kept on file for possible participation in future studies.

Results

Reliabilities. The primary assessment measures in this experiment were the test-retest reliability coefficients for the two item versions for each task format. In addition, we tested the difference between each item version's reliability to ascertain if one version or the other of

---

[2] Owing to the priority placed on estimating reliability and the size of the study, it was not possible to randomly cross question version (e.g., YN or Y0) with question content, so that differences between response times for different versions of the same basic format can be due to either version differences or content differences.

a task format produced higher reliabilities. Table 5 shows the reliabilities obtained for each of the four question sets under the YN and YO versions of this task format. All the reliability coefficients were significant at the p< .01 level or better. Results of the paired t-tests demonstrated that no reliable difference existed between the two versions. The reliabilities of the two versions of the ranking task are displayed in Table 6. All the reliabilities were at least significant at the p< .01 level except one, and that is significant at p< .05 level. The t-test outcomes indicated that the reliabilities for one of the questions differed significantly. Being required to move the ranks to the text might be a slightly more reliable procedure than moving the text to the ranks.

Table 7 depicts the reliabilities for the allocation task. Five of the reliabilities were significant at the p< .01 level and three are significant at the p< .05 level. Results of the t-test analyses showed that three of the four contrasts between the AP and AM versions were significant. Two of the contrasts were in favor of the AP version; whereas one favored the AM version. Overall, however, the balance seems to tip in favor of the AP version (overall means for the AP and AM versions are .729 and .680, respectively). Allocating percentages among various entities appeared a bit more reliable than allocating fixed amounts, such as money or time.

Regardless of whether respondents used a slider bar with a 0 to 10 scale or a slider bar with a 0 to 100 scale, mood assessments were equally reliable, as shown in Table 8. The reliability coefficients from both scales were significant at the p< .05 level.

TABLE 5. A CONTRAST BETWEEN YN AND YO RELIABILITY COEFFICIENTS

Question	YN Version Relib. Coef.	YO Version Relib. Coef.	Paired t-test	Signif of t
1	.755	.716	.265	n.s.
2	.862	.871	.136	n.s.
3	.920	.836	1.201	n.s.
4	.790	.716	.734	n.s.

TABLE 6. A CONTRAST BETWEEN RR AND RL RELIABILITY COEFFICIENTS

Question	RR Version Relib. Coef.	RL Version Relib. Coef.	Paired t-test	Signif of t
1	.686	.794	1.781	n.s.
2	.693	.856	4.702	.001
3	.818	.799	.209	n.s.
4	.743	.614	1.161	n.s.

TABLE 7. A CONTRAST BETWEEN AP AND AM RELIABILITY COEFFICIENTS

Question	AP Version Relib. Coef.	AM Version Relib. Coef.	Paired t-test	Signif of t
1	.790	.716	2.364	.038
2	.671	.744	3.702	.003
3	.843	.607	3.561	.004
4	.612	.654	1.234	n.s.

TABLE 8. A CONTRAST BETWEEN MOOD 1 AND 2 RELIABILITY COEFFICIENTS

Mood 1	Mood 2	Paired t-test	Signif of t
.549	.607	.542	n.s.

Response times. The transformed completion scores were submitted to a version (2) X order (2) within-subjects ANOVA, where order referred to whether Block 1 was followed by Block 2 or Block 2 was followed by Block 1. As expected, respondents took longer to complete the YN version than the YO version (means are 5.27 and 4.95, respectively; $F(1,10) = 19.10$, p<.001). We assumed that individuals took the time to consider each alternative when forced to respond yes or no to each question in contrast to considering only those that were true, as in the YO version. None of the other main effects for the other task formats attained acceptable significance levels. There were two significant version-by-order interactions: for the ranking task and the mood rating. Analysis of the order effect was performed as a check on the counterbalancing. We assume the two significant interaction were due to the imbalance created by the two "no-shows" in trial 2 and that the interactions were therefore artifacts of this condition.

Mood. The mood measures taken at the end of each Block were summed and submitted to a version (2) X order (2) within-subjects ANOVA. To compare the 1-10 scale measure with the 1-100 scale measure the latter scores were divided by 10. ANOVA revealed no main effects and a marginally significant interaction ($F(1,10) = 4.86$, p = .052). The scores for the scale version were about the same when Block 1 was followed by Block 2. When Block 2 was followed by Block 1, however, the mood of the respondents was higher when using the 1-100 scale. Respondents were either a little more positive when seeing the questions in Block 2 or using the longer scale first inflated the mood scores slightly. We cannot discriminate between these two hypotheses as the design was not set up to do so.

Discussion

As expected, when respondents were required to do some processing and consider each question for either a Yes or a No response, it took them longer to respond. There was an

indication, although not significant in this pilot study, that the longer processing time was bearing fruit in terms of higher reliabilities. There is one indication in the ranking task that moving the ranking tiles to the statements to be ranked produced more reliable data. There was no difference in the time it took to do the two ranking versions, so we assume moving the ranks to the statements represents a slightly more appealing or realistic procedure for the respondents. In two out of three significant results, allocating a percent produced more reliable data than allocating a fixed quantity like money or time. Prior to the experiment we were unsure if respondents would find it less cognitively demanding to allocate percents or a fixed quantity. At this point, the gauge has moved slightly in favor of allocating percent as less cognitively demanding, but the issue will require more investigation than provided by this pilot experiment. Respondents appeared to be equal in mood after Block 1 and Block 2. Counter to findings in the literature, the 1-10 scale length did not produce more reliable responses than the 1-100 scale length. Perhaps the mechanization of the slider bars in the GUI helped equalize the reliability of the two scales.

Overall, simply by programming such enhanced survey GUI question formats and employing them to collect data from actual respondents, we have demonstrated that respondents can be helped to provide responses that are faster, more complete, and less error-ridden.

<center>References</center>

Anderson, R. E. & Magnan, S. (1995). The Questionnaire Programming Language (QPL): An overview with examples of call management. <u>Social Science Computer Review</u> , 13:3:291-303.

Anderson, S. E., & Gansneder, B. M. (1995). Using electronic mail surveys and computer-monitored data for studying computer-mediated communication systems. <u>Social Science Computer Review, 13,</u> 3-46.

Barron, A. E., Tompkins, B., & Tai, D. (1996). Design guidelines for the World Wide Web. <u>Journal of Interactive Instruction Development, 8,</u> 13-17.

Blau, P. M. (1964). <u>Exchange and power in social life</u>. New York: Wiley.

Buetow, S. A., Douglas, R. M., Harris, P., & McCulloch, C. (1996). Computer-assisted personal interviews: Development and experience of an approach in Australian general practice. <u>Social Science Computer Review, 14,</u>:205-212.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., & Miller, P. (1987). An experimental comparison of telephone and personal health surveys. <u>Vital and Health Statistics</u>, Series 2, No. 106, Public Health Service (PHS), 87-1380.

Carpenter, E.H. (1975). Personalizing mail surveys: A replication and reassessment. <u>Public Opinion Quarterly, 36</u>, 254-257.

Catania, J. A., Binson, D., Canchola, J., Pollack, L. M., Hauck, W. & Coates, T. J. (1996). Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. <u>Public Opinion Quarterly, 60,</u>345-375.

Comber, T. (1995). Building usable web pages: An HCI perspective. http://www.scu.edu.au/ausweb95/papers/hypertext/comber/ Southern Cross University.

Couper, M. P., & Burt, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). <u>Social Science Computer Review, 12,</u> 38-54.

Couper, M. P., & Rowe, B. (1996). Evaluation of a computer-assisted self-interview component in a computer-assisted personal interview survey. <u>Public Opinion Quarterly, 60,</u> 89-105.

<center>35</center>

de Leeuw, E. D. & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R.M. Groves, P.P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), <u>Telephone survey methodology</u>. New York: Wiley.

Dillman, D.A., Carlson, J. E., & Lassey, W. R. (1978). <u>Absentee landowners and soil erosion control on Palouse farms</u>. College of Agriculture Research Center Circular No. 607. Pullman,WA: Washington State University.

Dillman, D. A. (1983). Mail and other self-administered questionnaires. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), <u>Handbook of survey research</u>.. New York: Academic Press.

Dillman, D. A., & Frey, J. H. (1974). Contribution of personalization to mail questionnaire response as an element of a previously tested method. <u>Journal of Applied Psychology, 59,</u> 296-301.

Groves, R. M., & Kahn, R. L. (1979). <u>Survey by telephone: A national comparison with personal interviews</u>. New York: Academic Press.

Groves, R. M., & Mathiowetz, N. (1984). Computer-assisted telephone interviewing: Effects on interviewers and respondents. <u>Public Opinion Quarterly, 48,</u> 356-369.

Helander, M. (Ed.). (1988). <u>Handbook of Human-Computer Interaction</u>. New York: Elsevier Science Publishing Company.

Herzog, A. R., & Rodgers, W. L. (1988). Interviewing older adults: Mode comparison using data from a face-to-face and a telephone resurvey. <u>Public Opinion Quarterly, 52,</u> 84-99.

Herzog, A. R., Rodgers, W. L., & Kulka, R. A. (1983). Interviewing older adults: A comparison of telephone and face-to-face modalities. <u>Public Opinion Quarterly, 47,</u> 405-418.

House, J. S., Gerber, W., & McMichael, A. J. (1977). Increasing mail questionnaire response: A controlled replication and extension. <u>Public Opinion Quarterly, 41,</u> 95-99.

Johnston, J., & Walton, C. (1995). Reducing response effects for sensitive questions: A computer-assisted self-interview with audio. <u>Social Science Computer Review, 13,</u> 304-319.

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response styles in telephone and household interviewing: A field experiment. <u>Public Opinion Quarterly, 44,</u> 210-222.

Kiesler, S., & Sproull, L. S. (1986). Response effects in electronic surveys. <u>Public Opinion Quarterly, 50,</u> 402-413.

Koltringer, R. (1995). Measurement quality in Austrian personal interview surveys. In Saris, W. E., & Munnich, A. (Eds.), The Multitrait-Multimethod Approach to Evaluate Measurement Instruments, Chapter 12. Budapest, Hungary: Eotvos University Press.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly, 51, 201-219.

Krysan, M., Schuman, H., Scott, L. J. & Beatty, P. (1994). Response rates and response content in mail versus face-to-face surveys. Public Opinion Quarterly, 58, 381-399.

Martin, C. I., & Nagao, D. H. (1989). Some effects of computerized interviewing on job application responses. Journal of Applied Psychology, 74, 72-80.

Nevin, J. R. & Ford, N. M. (1976). Effects of a deadline and a veiled threat on mail survey responses. Journal of Applied Psychology, 61, 116-118.

Nicholls II, W. L. (1988). Computer-assisted telephone interviewing: A general introduction. In R. M. Groves, P. P. Biemer, L.E. Lyberg, J. T. Massey, W. L. Nicholls II, and J. Waksberg (Eds.), Telephone survey methodology. New York: Wiley.

Nielsen, J. (1996). http://www.sun.com/sun-on-net/uidesign/

Parker, L. (1992). Collecting data the email way. Training and Development, 46, 52-54.

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1992). Quality of survey measures: A structural modeling approach. Journal of Official Statistics, 8, 251-275.

Rodman, R. D., & Williams, J. O. (1996). Hardware and software for telephoning voters en masse. Social Science Computer Review, 14,190-196.

Rossi, P. H., Wright, J. D., & Anderson, A. B. (1983). Sample surveys: History, current practice, and future prospects. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of survey research. New York: Academic Press.

Saris, W. E. (1991). Computer-Assisted Interviewing. Newbury Park, CA: Sage.

Scherpenzeel, A. C., & Saris, W. E. (1993). The evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies. Bulletin de Methodologie Sociologique, 39, 3-19.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. Sociological Methods & Research, 25, 341-383.

Sheatsley, P. B. (1983). Questionnaire construction and item writing. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of survey research.. New York: Academic Press.

Sykes, W., & Collins, M. (1988). Effects of mode of interview: Experiments in the UK. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J.T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), Telephone survey methodology. New York: Wiley.

Sproull, L. (1986). Using electronic mail for data collection in organizational research. Academy of Management Journal, 29, 159-169.

Thibaut, J. W., Kelly, H. H. (1959). The social psychology of groups. New York: Wiley.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. Public Opinion Quarterly, 60, 275-304.

Walsh, J. P., Kiesler, S., Sproull, L. S., & Hesse, B.W. (1992). Self-selected and randomly selected respondents in a computer network survey. Public Opinion Quarterly, 56, 241-244.

APPENDIX A

PROPOSED EXPERIMENTAL PROGRAM

The two pilot experiments provided a great deal of information that can be useful in planning and designing the future research. The pilot designs provided a good prototype for designs and methods to examine various issues such as layout, presentation, and sequencing where two versions or a presence/absence are to be contrasted in the same survey. The pilot experiments also defined a number of dependent measures that can be employed in subsequent research designs. In short, the two pilot experiments provided a basis for a subsequent stage of research.

In this appendix, we discuss several candidate areas for empirical research and outline an experimental program to conduct the research. These areas were identified through our analysis of survey technologies presented in our report. Our analysis indicates that although many aspects of survey methodology will generalize to the domain of computer-administered surveys over a computer network, a number of important issues still remain. Several of these issues are related to the computer's inherent abilities–the very abilities that will allow the design of active, that is, assisted-by-computer adaptive surveys.

Below we describe controlled experiments to examine these issues and procedures. Thus we delineate four sets of issues that we consider most critical to investigate: (1) cognitive features that refer to issues of animation, graphics, and text formatting; (2) layout, presentation, and sequencing issues that include variables such as single-item vs. multiple-item presentation, vertically vs. horizontally presented scales, numbered vs. verbally labeled scales, options to stop and restart the survey after a break, procedures to develop a modularized survey instrument, and procedures to launch automatic surveys; (3) confidentiality issues that address how best to elicit trust and assure anonymity and confidentiality convincingly in respondents; and (4) assessment of sensitive issues.

Research Issues

Cognitive Features

Collectively we refer to the issues of animation, graphics, and text formatting as cognitive features, and this is the first area we propose for future empirical research. We use the term cognitive features because these issues deal primarily with attracting and holding attention, aiding information processing by providing a rich medium (for example, instructions and aids), and increasing motivation by providing means to augment interest and give encouragements.

Surveys can be conducted in at least two different modes over a computer network: static and active. Transmitting what essentially is a "pencil-and-paper" mail survey over a

39

computer network is an example of a static survey. In this case, a survey is transmitted in its entirety, a few items at a time or item by item, to respondents' platforms. The respondents record their responses using whatever mechanisms are provided and then click a button to transmit their responses back to the host computer. The survey cannot be altered, modified, or adapted to the respondent except through the use of screening questions and skip (question) sequences. Using skip sequences is always risky as respondents often become confused and respond incorrectly or inappropriately (Dillman, 1983).

In the active mode, surveys can be prepared along with "intelligent" software running on the surveyor's computer to monitor the respondent's responses and modify or adapt the survey to the respondent. For example, the order in which questions are presented can be altered, the wording of questions can be modified or elaborated to accommodate reading skill or knowledge level, or a help function can be provided. Various levels of animation are possible throughout the survey to keep respondents interested, guide them through the survey, provide various response modes, or actually be part of the stimulus the survey is evaluating, judging, or rating.

Clearly, active adaptive surveys are the wave of the future. To create efficient and effective surveys we must know how to use effectively the various capabilities of the computer and network. Dillman (1983) underscored the requirement that self-administered surveys must appear attractive and inviting. He suggested that an appropriate illustration or graphic be placed on the front cover of the survey booklet to catch the potential respondent's eye and engender interest. But illustrations, graphics, and animation need not be limited to the front of the survey. The pilot studies reported in the main body of this report indicated that appropriate formatting and GUI can reduce effort and help produce error-free data.

Employing a computer to administer a survey, either directly or over a network, opens a plethora of possibilities to embellish the survey, such as different colored text or borders around text, flashing words or symbols, color illustrations or graphics, animated pointers, dots, or arrows, or animated illustrations or graphics. Granted many of these additions would be eye-catching and some interesting. But, can a survey look too cute, pretty, or slick? If inappropriately used, can graphics and animation undermine trust or be a put-off to potential respondents? Perhaps, in the end, simple is best. Obviously, these issues need to be examined empirically in light of what is already known from existing theory and research. Guidelines are needed to help survey developers maximize the effects of graphics or animation on response rate, compliance, trust, and the like. Analysis and/or empirical research might determine that some forms of embellishment are more distracting than others or are more appropriate in one kind of situation than another. As suggested earlier in this report, guidelines derived from human-computer interaction studies will be applicable here.

Layout, Presentation, and Sequencing Issues

This second area is concerned with layout or presentation issues and sequencing or "controlling the flow" issues. Layout and presentation problems include variables such as single-item vs. multiple-item presentations, vertically vs. horizontally presented scales, numbered vs. verbally labeled scales. Sequencing issues include variables such as options to stop and restart the survey after a break, procedures to develop a modularized survey instrument, and procedures to launch automatic surveys. Viable candidates for layout and presentation research include some combination of: a time estimate in minutes/hours, clock or time bar that is dynamically updated, total number of questions, or a question counter or question bar that decreases with each completed item. Sequencing issues pose technical as well as research issues; however, implementing surveys with pauses or breaks and assessing their effect on response data are within existing capabilities.

With a paper-and-pencil, self-administered mail survey, the stimulus is present in front of the individuals. They can see the number of pages, observe the number of questions, and perhaps even gauge the difficulty of completing the questions. In short, they can make an estimate of how long they think the survey will take to complete. Based on this estimate, individuals can decide to start the questionnaire, delay starting until a more convenient time, or choose not to do the survey at all. When a survey is presented by a computer network, it may not be possible to scroll through the questionnaire to gauge how long it will take to complete. Rather, respondents will only see one or perhaps a few questions at a time. What effect might this serial presentation of questions of a survey instrument of unknown length have on response rate? More than likely, the author of a survey presented over a network will want to provide some kind of estimate of how long the questionnaire will take to complete. What form should this estimate take to maximize response rate? An inappropriately chosen time or length estimate may get respondents started, only to have them abandon the questionnaire before completion.

The ability to create an active and adaptive survey also provides the ability to conduct studies not before possible, e.g., distributed surveys. Assume, for example, that the intention is to sample several units in the military (e.g., companies, brigades, or battalions). Suppose the survey protocol calls for ten people from each unit to be surveyed. If each unit maintained an electronic data base of personnel and the data bases were connected to a computer network, a likely possibility, then a JAVA-like applet could be written to interrogate each data base for specific statistics. Based on these statistics the applet would select the most appropriate representative or random sample to be sent the survey. The applet would then route the survey to the selected individuals and administer it. Detailed information about the specific units does not have to be known centrally and, once launched, the survey would run automatically. The partitioning of a survey may also require some mechanism to pause or interrupt work on it and then resume it later. Respondents working on very long or complex surveys may need some way to pause the survey until they can continue work. By what mechanisms could this be done, what is the best way, what impact would this have on the response data?

As another illustration, consider that frequently in the military, government, or large corporations the individual receiving a survey may not be the individual actually

completing it. That is, the individual receiving the survey may wish to parcel out different parts of the survey to one or more appropriate staff members for completion. An active survey could be developed to accommodate such circumstances. First, the survey author must develop a modularized survey instrument, which is sent to the initial respondent. The initial respondent is given the opportunity to route each module to a particular staff member or to respond to it personally. The applet administering the survey would distribute the modules to the specified people for completion. Once each module is completed, the applet reassembles the survey for the initial respondent's inspection and editing. When satisfied, the initial respondent releases the survey back to the host computer. The possibility of such a survey process evokes a host of questions. What effect would this process have on response behavior (of each individual involved) and on data quality? Is speed of completion necessarily governed by the slowest link? What does it mean to have a multi-respondent survey? Which issues and questions are best addressed by such a survey process and which are not? The preceding is only a small set of the potential questions that would be raised by such a distributed survey process.

Confidentiality

An important variable impacting compliance and the willingness of individuals to participate in a survey is the confidentiality their responses will be afforded. Many survey researchers ardently believe that people will not participate in a survey, or will not respond honestly, unless they are assured their responses will be confidential (Singer, Thurn, & Miller, 1995; Singer, 1978). Hawkins (1977) noted that the nonresponse rate has climbed from 15 percent to 30 percent in the past 20 years for most survey research groups. Brooks and Bailar (1978) reported that an increasing proportion of noninterviews is accounted for by refusals. Not surprisingly, in the era of decreased confidence in government and corporate integrity, individual confidence in confidentiality of responses has declined. A National Academy of Science survey indicated that only five percent of the respondents believed that census records were truly confidential, whereas 80 percent reported that they did not believe the census data were confidential or that confidentiality could be maintained if other agencies of the government really wanted to obtain the records (NAS, 1979). There are not many studies examining confidentiality and refusal to participate in a survey (Singer et al., 1995, noted there were none before 1975), but in one study it was reported that the presence or absence of a confidentiality statement and the strength of that statement had a consistent effect on refusal rate (Martin, 1983). Singer et al. (1995) noted that their findings indicated a link between assurances of confidentiality and response quality, but only for responses concerning sensitive issues. Thus, it is understandable that if individuals' confidence in confidentiality has been undermined for whatever reason, they will be less willing to participate–particularly if the information sought is perceived as potentially damaging or embarrassing. Clearly the assurance of confidentiality is a key ingredient in soliciting sensitive information.

Surveys to be administered over a computer network are like mail surveys in that participation and compliance are solicited through written introductions. The surveyor attempts to elicit trust and almost invariably promises that the responses collected will be held

in confidence. The potential respondent must come to trust the surveyor. Several factors may abet this trust such as organizational affiliation, perceived status, and affiliation with the surveyor. Surveys that are administered through universities or known survey organization may be perceived as more scientific, professional, and unbiased and hence more trustworthy. Similarly surveyors presented as professor, doctor, or pastor or someone from the respondent's school, church group, or professional organization may also engender confidence and trust.

A crucial factor employed by surveyors to guarantee confidentiality of responses is to offer the respondents anonymity. Except for mail surveys that can be completed anonymously and mailed back without attribution, the respondents to other modes of administration must trust the surveyor that their responses will be recorded without attribution. Administration of network surveys is no exception. Currently all correspondence over computer networks bears the affiliations of the parties involved. Although new technology may make it possible to send correspondence over computer networks anonymously in some cases, most respondents will have to be persuaded to trust the surveyor as to anonymity and confidentiality. This task has not been made any easier by reports in the media of improprieties and invasions of privacy on the Internet.

Accordingly, we must determine how best to write and present introductions over a computer network to convincingly convey the promise of anonymity and confidentiality to potential respondents. Another question will be how best to elicit trust in the respondents. Research should also be carried out to assess varying beliefs in anonymity and confidentiality of surveys completed over the network and correlate these with data reliability and validity.

Assessment of Sensitive Issues

An important reason for the growth of self-administered paper and pencil surveys and particularly CASI is to tackle the difficult problem of surveying sensitive issues. By sensitive issues we mean those that may embarrass the responder (e.g., questions about sexual practices, the contraction of venereal diseases) or inquire about extralegal practices (e.g., drug usage, welfare cheating). The findings of one study showed it was not the computerization of the survey per se, but the self-administered aspect, whether by pencil and paper or computer, that had a clear impact on reporting–particularly the reporting of sexual behavior (Tourangeau & Smith, 1996). Moreover, studies have reported that respondents felt that surveys conducted by computer were more important and objective (Tourangeau & Smith, 1996), that self administration reduced fears of embarrassment and increased candor (Ferriter, 1993; Plutchik & Karasu, 1991), including extremes of responses (Ferriter, 1993; Thornberry, Rowe, & Biggar, 1991). There are also reports of a reduction both in underreporting (Duffy & Wateron, 1984) and bias toward socially desirable responses (Ferriter, 1993). Self-assessment and

computerization thus appear to combine the best of both worlds (Johnston & Walton, 1995). CASI and its close relative audio-CASI (ACASI), which circumvents the problem of reading literacy, have been reported to yield the highest reported incidents of oral and anal sex and

drug usage (Tourangeau & Smith, 1996).

Although both CASI and ACASI do not require the respondents to report answers to an interviewer, there is an interviewer present who has introduced the survey, appealed for trust, and promised confidentiality. Individuals responding to a survey administered over a computer network may not have the same feelings of confidentiality and, lacking true anonymity, may be reluctant to answer sensitive questions. Some studies examining computer-aided interviews have not found increased response candor (e.g., Skinner, Allen, McIntosh, & Palmer, 1985). Lynch (1996), investigating the emotional and sensitive area of rape, did not feel that the use of CASI would yield categorically better responses than other survey modes. His findings indicated that the design of the survey may be most important for such emotionally laden issues.

<u>Experimental Designs</u>

Issues from the first two research areas (cognitive features and layout presentation, and sequencing) are amenable to systematic investigation employing the research design developed for the pilot experiments reported in the body of this report. One to three independent variables representing operationalization of issues from these two areas could be manipulated over two levels (presence and absence–to keep matters simple initially) and completely crossed to produce up to eight experimental conditions. For example, we could cross the presence or absence of an on-demand help function, graphical illustrations for examples, and animation as an attention-getting and motivational device. Or we can take two of the previous factors and cross them with the number of items sent to the screen (one, small group, or whole survey). In most situations, response biases will be masked by the variability associated with respondents' knowledge or opinions or attitudes on question content. To make response biases apparent, then, a large number of observations will be needed to reduce the error variance for statistical testing. This could be achieved by using large samples of respondents, but that may be prohibitively expensive and is not administratively flexible enough to allow the pursuit of many questions seriatim. A better alternative is to administer a great many items to smaller groups of experimental and control subjects and to take advantage of the statistical power gained thereby. In statistical terms this is a split-plot design used to compare between groups (e.g., computer vs. control presentation) using repeated measures (subjects by trials).

A number of dependent variables were delineated in the pilot experiments and, depending on the issues involved, a set of appropriate measure can be selected to assess the effects of the independent factors. Given the general nature of the experiments involving cognitive factors and layout, presentation, and sequencing issues, college students would

suffice as subjects. However, as demonstrated in pilot Experiment 2, nonstudent populations are also desirable.

Some of the issues in the confidentiality and sensitive issue research areas might be amenable to a design such as that described above, but more likely a parallel survey design will be called for. In the latter design, between-group comparisons could be used to test differences between computer-based presentations and traditional paper-and-pencil presentations, between computer-based and aural presentations, or among several different computer-based presentation formats. For example, a survey can be devised to gain information as to the respondents' use of several illicit drugs. The introduction to half the survey respondents could include a carefully worded statement assuring anonymity and confidentiality of responses, whereas the other half could have no such statement or a one-line statement to that effect. The survey could be prepared to be administered in two modes: over a computer network and as a traditional mail or telephone survey. Comparison across the two modes of administration would yield information on the effect of computer network administration on sensitive issues and how the promise of anonymity and confidentiality affected response rate and quality. Given the nature of such parallel survey designs, they can be launched from anywhere there is access to the Internet or other suitable computer network and a mail box.

References—Appendix

45

Brooks, C. A., & Bailar, B. A. (1978). An error profile: Employment as measured by the current population survey. Statistical Policy Working Paper 3. Washington D.C.: U.S. Department of Commerce.

Dillman, D. A. (1983). Mail and other self-administered questionnaires. In P.H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of survey research. New York: Academic Press.

Duffy, J. C., & Waterton, J. J. (1984). Under-reporting of alcohol consumption in sample surveys: The effect of computer interviewing in fieldwork. British Journal of Addiction, 79, 303-308.

Ferriter, M. (1993). Computer-aided interviewing in psychiatric social work. Computers in Human Services, 9, 59-66.

Hawkins, D. F. (1977). Nonresponse in Detroit area study surveys: A ten-year analysis. Working Paper in Methodology No. 8. Chapel Hill, N.C.: Institute for Research in Social Science.

Johnston, J., & Walton, C. (1995). Reducing response effects for sensitive questions: A computer-assisted self-interview with audio. Social Science Computer Review, 13, 304-319.

Lynch, J. P. (1996). The polls-review: Clarifying divergent estimates of rape from two national surveys. Public Opinion Quarterly, 60:410-430.

Martin, E. (1983). Surveys as social indicators: Problems in monitoring trends. In P.H. Rossi, J.D. Wright, and A.B. Anderson (Eds.), Handbook of survey research. NewYork: Academic Press.

National Academy of Sciences (NAS) (1979). Privacy and confidentiality as factors in survey response. Washington D.C.: Author.

Plutchik, R., & Karasu, T. B. (1991). Computers in psychotherapy: An overview. Computers in Human Behavior, 7, 33-44.

Singer, E. (1978). The effect of informed consent procedures on respondents' reactions to surveys. Journal of Consumer Research, 5, 49-57.

Singer, E., von Thurn, D. R., & Miller, E. R. (1995). Confidentiality assurances and response: A quantitative view of the experimental literature. Public Opinion Quarterly, 59, 66-77.

Skinner, H. A., Allen, B. A., McIntosh, M. C., & Palmer, W. H. (1985). Lifestyle assessments: Applying microcomputers in family practice. <u>British Journal of Clinical Research and Education, 290</u>, 212-214.

Thornberry, O., Rowe, B., & Biggar, R. (1991). Use of CAPI with the National Health Insurance Survey. <u>Bulletin de Methodologie Sociologique, 30,</u> 27.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. <u>Public Opinion Quarterly, 60</u>, 275-304.